Evolutionary dynamics in set structured populations

Corina E. Tarnita[a], Tibor Antal[a], Hisashi Ohtsuki[b], and Martin A. Nowak[a,1]

[a]Program for Evolutionary Dynamics, Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; and [b]Department of Value and Decision Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan

Communicated by Robert May, University of Oxford, Oxford, United Kingdom, April 2, 2009 (received for review February 8, 2009)

Evolutionary dynamics are strongly affected by population structure. The outcome of an evolutionary process in a well-mixed population can be very different from that in a structured population. We introduce a powerful method to study dynamical population structure: evolutionary set theory. The individuals of a population are distributed over sets. Individuals interact with others who are in the same set. Any 2 individuals can have several sets in common. Some sets can be empty, whereas others have many members. Interactions occur in terms of an evolutionary game. The payoff of the game is interpreted as fitness. Both the strategy and the set memberships change under evolutionary updating. Therefore, the population structure itself is a consequence of evolutionary dynamics. We construct a general mathematical approach for studying any evolutionary game in set structured populations. As a particular example, we study the evolution of cooperation and derive precise conditions for cooperators to be selected over defectors.

cooperation | game | social behavior | stochastic dynamics

Human society is organized into sets. We participate in activities or belong to institutions where we meet and interact with other people. Each person belongs to several sets. Such sets can be defined, for example, by working for a particular company, living in a specific location, going to certain restaurants, or holding memberships at clubs. There can be sets within sets. For example, the students of the same university have different majors, take different classes, and compete in different sports. These set memberships determine the structure of human society: they specify who meets whom, and they define the frequency and context of meetings between individuals.

We take a keen interest in the activities of other people and contemplate whether their success is correlated with belonging to particular sets. It is therefore natural to assume that we do not only imitate the behavior of successful individuals, but also try to adopt their set memberships. Therefore, the cultural evolutionary dynamics of human society, which are based on imitation and learning, should include updating of strategic behavior and of set memberships. In the same way as successful strategies spawn imitators, successful sets attract more members. If we allow set associations to change, then the structure of the population itself is not static, but a consequence of evolutionary dynamics.

There have been many attempts to study the effect of population structure on evolutionary and ecological dynamics. These approaches include spatial models in ecology (1–8), viscous populations (9), spatial games (10–15), and games on graphs (16–19).

We see "evolutionary set theory" as a powerful method to study evolutionary dynamics in structured populations in the context where the population structure itself is a consequence of the evolutionary process. Our primary objective is to provide a model for the cultural evolutionary dynamics of human society, but our framework is applicable to genetic evolution of animal populations. For animals, sets can denote living at certain locations or foraging at particular places. Any one individual can belong to several sets. Offspring might inherit the set memberships of their parents. Our model could also be useful for studying dispersal behavior of animals (20, 21).

Let us consider a population of N individuals distributed over M sets (Fig. 1). Individuals interact with others who belong to the same set. If 2 individuals have several sets in common, they interact several times. Interactions lead to payoff from an evolutionary game.

The payoff of the game is interpreted as fitness (22–26). We can consider any evolutionary game, but at first we study the evolution of cooperation. There are 2 strategies: cooperators, C, and defectors, D. Cooperators pay a cost, c, for the other person to receive a benefit, b. Defectors pay no cost and provide no benefit. The resulting payoff matrix represents a simplified Prisoner's Dilemma. The crucial parameter is the benefit-to-cost ratio, b/c. In a well-mixed population, where any 2 individuals interact with equal likelihood, cooperators would be outcompeted by defectors. The key question is whether dynamics on sets can induce a population structure that allows evolution of cooperation.

Individuals update stochastically in discrete time steps. Payoff determines fitness. Successful individuals are more likely to be imitated by others. An imitator picks another individual at random, but proportional to payoff, and adopts his strategy and set associations. Thus, both the strategy and the set memberships are subject to evolutionary updating. Evolutionary set theory is a dynamical graph theory: who interacts with whom changes during the evolutionary process (Fig. 2). For mathematical convenience we consider evolutionary game dynamics in a Wright–Fisher process with constant population size (27). A frequency-dependent Moran process (28) or a pairwise comparison process (29), which is more realistic for imitation dynamics among humans, give very similar results, but some aspects of the calculations become more complicated.

The inheritance of the set memberships occurs with mutation rate v: with probability $1 - v$, the imitator adopts the parental set memberships, but with probability v a random sample of new sets is chosen. Strategies are inherited subject to a mutation rate, u. Therefore, we have 2 types of mutation rates: a set mutation rate, v, and a strategy mutation rate, u. In the context of cultural evolution, our mutation rates can also be seen as "exploration rates": occasionally, we explore new strategies and new sets.

We study the mutation-selection balance of cooperators versus defectors in a population of size N distributed over M sets. In the supporting information (SI) *Appendix*, we show that cooperators are more abundant than defectors (for weak selection and large population size) if $b/c > (z - h)/(g - h)$. The term z is the average number of sets 2 randomly chosen individuals have in common. For g we pick 2 random, distinct, individuals in each state; whether they have the same strategy, we add their number of sets to the average, otherwise we add 0; g is the average of this average over the stationary distribution. For understanding h we must pick 3 individuals at random: then h is the average number of sets the first 2 individuals have in common

Author contributions: C.E.T., T.A., H.O., and M.A.N. performed research and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0903019106/DCSupplemental.

EVOLUTION

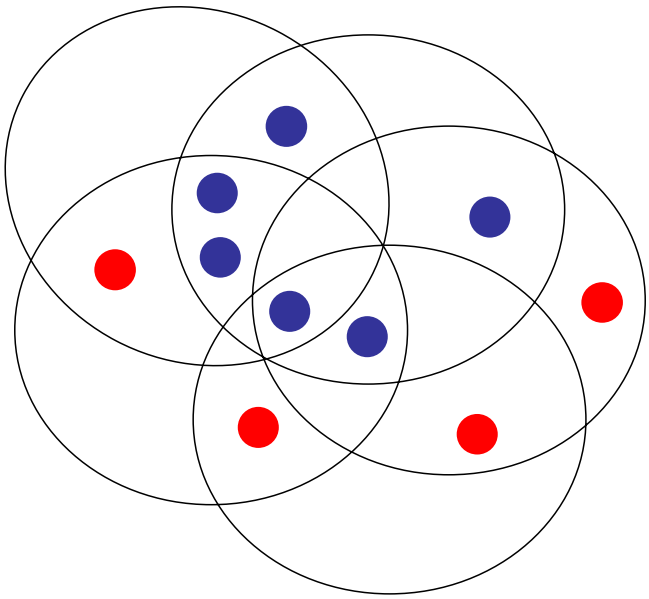


Fig. 1. Evolutionary set theory offers a powerful approach for modeling population structure. The individuals of a population are distributed over sets. Individuals can belong to several sets. Some sets might contain many individuals, whereas others are empty. Individuals interact with others who share the same sets. These interactions lead to payoff from evolutionary games. Both the strategy (= behavior in the game) and the set memberships of individuals are updated proportional to payoff. Sets that contain successful individuals attract more members. In this example, there are 2 strategies (indicated by red and blue). The population size is $n = 10$, and there are $M = 5$ sets.

given that there is a nonzero contribution to the average only if the latter 2 share the same strategy. We prove in the *SI Appendix* that for the limit of weak selection these 3 quantities can be calculated from the stationary distribution of the stochastic process under neutral drift. Our calculation uses arguments from perturbation theory and coalescent theory that were first developed for studying games in phenotype space (30).

Analytical calculations are possible if we assume that each individual belongs to K out of the M sets. We obtain exact results for any choice of population size and mutation rates (see *SI Appendix*). For large population size N and small strategy mutation rate, u, the critical benefit to cost ratio is given by

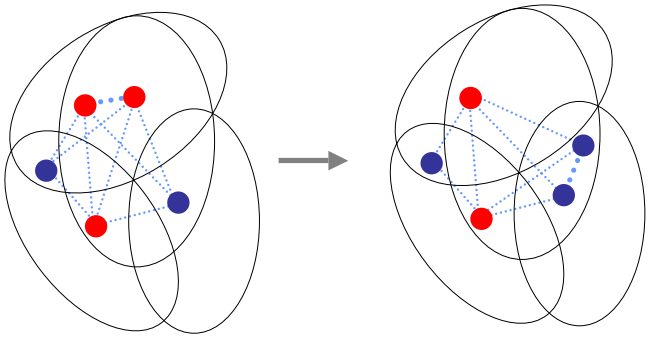


Fig. 2. Evolutionary set theory is a dynamical graph theory. The set memberships determine how individuals interact. If 2 individuals have several sets in common, then they interact several times. In this example, there are $N = 5$ individuals over $M = 4$ sets. Each individual belongs to $K = 2$ sets. The broken lines indicate the weighted interaction graph. The graph changes as one individual moves to new sets. The evolutionary updating of strategies and set memberships happens at the same timescale.

Table 1. The minimum benefit-to-cost ratio

K	L	K^*	Minimum b/c
1	1	1	2.17
2	1	2	3.13
2	2	2/9	1.42
3	1	3	4.29
3	2	5/4	2.30
3	3	1/12	1.24
4	1	4	5.80
4	2	64/21	4.35
4	3	19/21	2.09
4	4	1/21	1.18
5	1	5	7.94
5	2	605/126	7.44
5	3	45/14	4.58
5	4	5/6	2.02
5	5	5/126	1.16
6	1, 2	6	11.26
6	3	121/21	10.31
6	4	27/7	5.55
6	5	1	2.17
6	6	1/21	1.18
7	1, 2, 3, 4	7	17.23
7	5	16/3	8.87
7	6	19/12	2.72
7	7	1/12	1.24
8	1, 2, 3, 4, 5, 6	8	31.24
8	7	10/3	4.74
8	8	2/9	1.42
9	1, 2, 3, 4, 5, 6, 7, 8	9	104.7
9	9	1	2.17

Each individual belongs to K sets. Cooperation is triggered if the other person has at least L sets in common. For a fixed total number of sets, $M = 10$, we can vary K = 1,. . .,9 and L = 1,. . .,K. The effective number of set memberships, K^*, is given by Eq. **54** in the *SI Appendix*. For a finite population of $N = 100$, we calculate the minimum benefit-to-cost ratio that is needed for selection to favor cooperators over defectors. For given K, cooperators are favored by larger L, and the smallest b/c ratio is obtained by $L = K$. The global minimum occurs for $L = K = M/2$. Smaller values of K^* imply smaller values of the critical b/c. Note that various symmetries lead to identical K^* values.

$$\frac{b}{c} = \frac{K}{M-K}(\nu+2) + \frac{M}{M-K}\frac{\nu^2+3\nu+3}{\nu(\nu+2)}. \qquad [1]$$

Here, we introduced the parameter $\nu = 2N\nu$, which denotes an effective set mutation rate scaled with population size: it is twice the expected number of individuals who change their set memberships per evolutionary updating. Eq. **1** describes a U-shaped function of ν. If the set mutation rate is too small, then all individuals belong to the same sets. If the set mutation rate is too large, then the affiliation with individual sets does not persist long enough in time. In both cases, cooperators have difficulties to thrive. But there exists an intermediate optimum set mutation rate which, for $K \ll M$ is given by $\nu_{\text{opt}} = \sqrt{M/K}$. At this optimum value of ν, we find the smallest critical b/c ratio for cooperators to outcompete defectors. If each individual can only belong to a small number of all available sets, $K \ll M$, then the minimum ratio is given by

$$\left(\frac{b}{c}\right)_{\min} = 1 + 2\sqrt{\frac{K}{M}}. \qquad [2]$$

Smaller values of K and larger values of M facilitate the evolution of cooperation. For a given number of sets, M, it is best if individuals belong to $K = 1$ set. Larger values of K make it harder for cooperators to avoid exploitation by defectors. But the

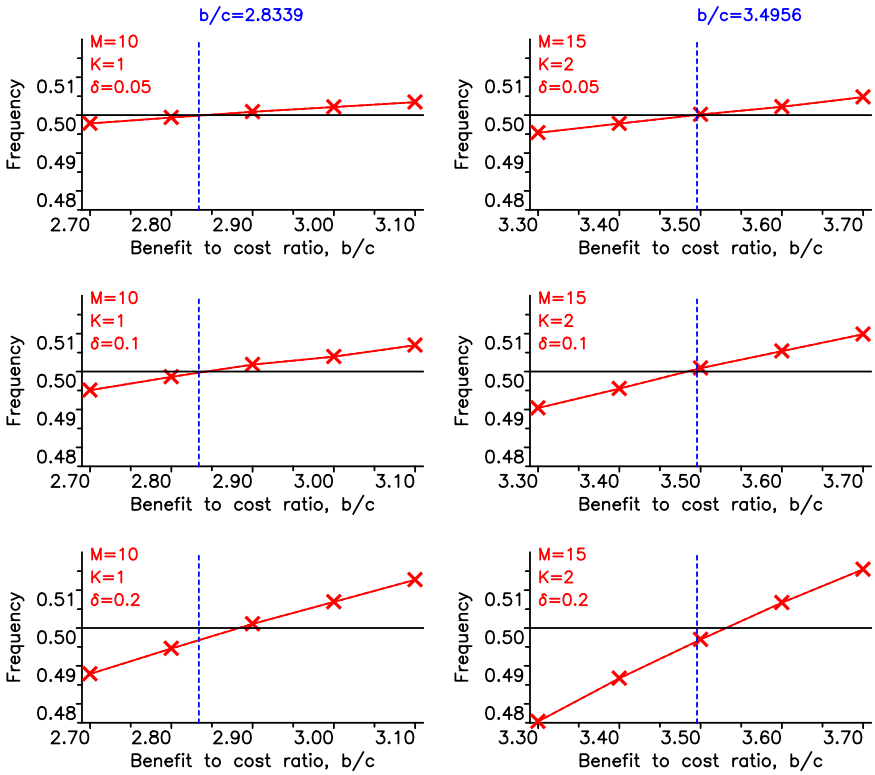


Fig. 3. Agreement between numerical simulations and the analytical theory. A population of size $N = 40$ is distributed over $M = 10$ or $M = 15$ sets. Individuals can belong to $K = 1$ or $K = 2$ sets. Each point indicates the frequency of cooperators averaged over 5×10^8 generations. Increasing the benefit-to-cost ratio, b/c, favors cooperators. At a certain b/c ratio cooperators become more abundant than defectors (intersection with the horizontal line). We study 3 different intensities of selection, $\delta = 0.05$, 0.1, and 0.2. There is excellent agreement between the numerical simulations and the analytical prediction for weak selection, which is indicated by the vertical blue line. Other parameters values are $L = 1$, $u = 0.002$, $v = 0.1$, $b = 1$, and c varies as indicated.

extension, which we will discuss now, changes this particular property of the model and makes it advantageous to be in more sets.

Let us generalize the model as follows: As before, defectors always defect, but now cooperators only cooperate, if they have a certain minimum number of sets, L, in common with the other person. If a cooperator meets another person in i sets, then the cooperator cooperates i times if $i \geq L$; otherwise cooperation is not triggered. $L = 1$ brings us back to the previous framework. Large values of L mean that cooperators are more selective in choosing with whom to cooperate. Interestingly, it turns out that this generalization leads to the same results as before, but K is replaced by an "effective number of set memberships," K^*, which does not need to be an integer and can even be <1 (see *SI Appendix*). In Table 1, we show K^* and the minimum b/c ratio for a fixed total number of sets, M, and for any possible values of K and L. For any given number of set memberships, K, larger values of L favor cooperators. We observe that belonging to more sets, $K > 1$, can facilitate evolution of cooperation, because for given M the smallest minimum b/c ratio is obtained for $K = L = M/2$.

In Fig. 3, we compare our analytical theory with numerical simulations of the mutation-selection process for various parameter choices and intensities of selection. We simulate evolutionary dynamics on sets and measure the frequency of cooperators averaged over many generations. Increasing the benefit-to-cost ratio favors cooperators, and above a critical value they become more abundant than defectors. The theory predicts this critical b/c ratio for the limit of weak selection. We observe that for decreasing selection intensity the numerical results converge to the theoretical prediction.

Our theory can be extended to any evolutionary game. Let A and B denote 2 strategies whose interaction is given by the payoff matrix $[(R, S),(T, P)]$. We find that selection in set structured populations favors A over B provided $\sigma R + S > T + \sigma P$. The value of σ is calculated in the *SI Appendix*. A well-mixed population is given by $\sigma = 1$. Larger values of σ signify increasing effects of population structure. We observe that σ is a one-humped function of the set mutation rate, ν. The optimum value of ν, which maximizes σ, is close to $\sqrt{M/K^*}$. For $K = L = 1$ the maximum value of σ grows as $\sqrt{M}$, but for $K = L = M/2$ the maximum value of σ grows exponentially with M. This demonstrates the power of sets.

Suppose A and B are 2 Nash equilibrium strategies in a coordination game, defined by $R > T$ and $P > S$. If $R + S < T + P$, then B is risk-dominant. If $R > P$ then A is Pareto efficient. The well-mixed population chooses risk dominance, but if σ is large enough, then the set structured population chooses the efficient equilibrium. Thus, evolutionary dynamics on sets can select efficient outcomes.

We have introduced a powerful method to study the effect of a dynamical population structure on evolutionary dynamics. We have explored the interaction of 2 types of strategies: unconditional defectors and cooperators who cooperate with others if they are in the same set or, more generally, if they have a certain number of sets in common. Such conditional cooperative behavior is supported by what psychologists call social identity theory (31). According to this idea people treat others more favorably, if they share some social categories (32). Moreover, people are motivated to establish positive characteristics for the groups with whom they identify. This implies that cooperation within sets is more likely than cooperation with individuals from other sets. Social identity theory suggests that preferential

cooperation with group members exists. Our results show that it can be adaptive if certain conditions hold. Our approach can also be used to study the dynamics of tag based cooperation (33–35): some of our sets could be seen as labels that help cooperators to identify each other.

In our theory, evolutionary updating includes both the strategic behavior and the set associations. Successful strategies leave more offspring, and successful sets attract more members. We derive an exact analytic theory to describe evolutionary dynamics on sets. This theory is in excellent agreement with numerical simulations. We calculate the minimum benefit-to-cost ratio that is needed for selection to favor cooperators over defectors. The mechanism for the evolution of cooperation (36) that is at work here is similar to spatial selection (10) or graph selection (17). The structure of the population allows cooperators to "cluster" in certain sets. These clusters of cooperators can prevail over defectors. The approach of evolutionary set theory can be applied to any evolutionary game or ecological interaction.

ACKNOWLEDGMENTS. This work was supported by the John Templeton Foundation, the National Science Foundation/National Institutes of Health joint program in mathematical biology (National Institutes of Health Grant R01GM078986), the Japan Society for the Promotion of Science, and J. Epstein.

1. MacArthur RH, Wilson EO (1967) *The Theory of Island Biogeography* (Princeton Univ Press, Princeton, NJ).
2. Levins R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull Entomol Soc Am* 15:237–240.
3. Levin SA, Paine RT (1974) Disturbance, patch formation, and community structure. *Proc Natl Acad Sci USA* 71:2744–2747.
4. Kareiva P (1987) Habitat fragmentation and the stability of predator-prey interactions. *Nature* 326:388–390.
5. Durrett R, Levin SA (1994) Stochastic spatial models: A user's guide to ecological applications. *Philos Trans R Soc London Ser B* 343:329–350.
6. Hassell MP, Comins HN, May RM (1994) Species coexistence and self-organizing spatial dynamics. *Nature* 370:290–292.
7. Tilman D, Kareiva P, eds (1997) *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (Princeton Univ Press, Princeton, NJ).
8. May RM (2006) Network structure and the biology of populations. *Trends Ecol Evol* 21:394–399.
9. Hamilton WD (1964) The genetical evolution of social behavior. *J Theor Biol* 7:1–16.
10. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359:826–829.
11. Killingback T, Doebeli M (1996) Spatial evolutionary game theory: Hawks and Doves revisited. *Proc R Soc London Ser B* 263:1135–1144.
12. Nakamaru M, Matsuda H, Iwasa Y (1997) The evolution of cooperation in a lattice structured population. *J Theor Biol* 184:65–81.
13. Szabó G, Töke C (1998) Evolutionary prisoner's dilemma game on a square lattice. *Phys Rev E* 58:69–73.
14. Kerr B, Riley MA, Feldman MW, Bohannan BJ (2002) Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* 418:171–174.
15. Hauert C, Doebeli M (2004) Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* 428:643–646.
16. Lieberman E, Hauert C, Nowak MA (2005) Evolutionary dynamics on graphs. *Nature* 433:312–316.
17. Ohtsuki H, Hauert C, Lieberman E, Nowak MA (2006) A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441:502–505.
18. Taylor PD, Day T, Wild G (2007) Evolution of cooperation in a finite homogeneous graph. *Nature* 447:469–472.
19. Santos FC, Santos MD, Pacheco JM (2008) Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454:213–216.
20. Hamilton WD, May RM (1977) Dispersal in stable habitats. *Nature* 269 578–581.
21. Taylor PD (1988) An inclusive fitness model for dispersal of offspring. *J Theor Biol* 130:363–378.
22. Maynard Smith J (1982) *Evolution and the Theory of Games* (Cambridge Univ Press, Cambridge, UK).
23. Colman AM (1995) *Game Theory and Its Applications in the Social and Biological Sciences* (Butterworth-Heinemann, Oxford, UK).
24. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, UK).
25. Cressman R (2003) *Evolutionary Dynamics and Extensive Form Games* (MIT Press, Cambridge, MA).
26. Nowak MA, Sigmund K (2004) Evolutionary dynamics of biological games. *Science* 303:793–799.
27. Imhof LA, Nowak MA (2006) Evolutionary game dynamics in a Wright-Fisher process. *J Math Biol* 52:667–681.
28. Nowak MA, A Sasaki, C Taylor, D Fudenberg (2004) Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428:646–650.
29. Traulsen A, Pacheco JM, Nowak MA (2007) Pairwise comparison and selection temperature in evolutionary game dynamics. *J Theor Biol* 246:522–529.
30. Antal T, Ohtsuki H, Wakeley J, Taylor PD, Nowak MA (2008) Evolution of cooperation by phenotypic similarity. *Proc Natl Acad Sci USA*, 10.1073/pnas.0902528106.
31. Tajfel H (1982) Social psychology of intergroup relations. *Annu Rev Psychol* 33:1–30.
32. Yamagishi T, Jin N, Kiyonari T (1999) Bounded generalized reciprocity. *Adv Group Process* 16:161–197.
33. Riolo RL, Cohen MD, Axelrod R (2001) Evolution of cooperation without reciprocity. *Nature* 418:441–443.
34. Traulsen A, Claussen JC (2004) Similarity based cooperation and spatial segregation. *Phys Rev E* 70:046128.
35. Jansen VA, van Baalen M (2006) Altruism through beard chromodynamics. *Nature* 440:663–666.
36. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314:1560–1563.

Supplementary Information for

Evolutionary Dynamics in Set Structured Populations

Corina E. Tarnita[1], Tibor Antal[1], Hisashi Ohtsuki[2], Martin A. Nowak[1]

[1] Program for Evolutionary Dynamics, Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2] Department of Value and Decision Science, Tokyo Institute of Technology, Tokyo, 152-8552, Japan

This 'Supplementary Information' has the following structure. In Section 1 ('Evolution of cooperation on sets') we discuss the basic model and derive a general expression for the critical benefit-to-cost ratio. In Section 2 ('Belonging to K sets') we calculate the critical benefit-to-cost ratio for the model where all individuals belong to exactly K sets and cooperators cooperate with all other individuals in the same set. In Section 3 ('Triggering cooperation ') we generalize the basic model to the situation where cooperators are only triggered to cooperate if the other individual has at least L sets in common. In Section 4 ('The minimum benefit-to-cost ratio') we calculate the optimum set mutation rate that minimizes the critical benefit-to-cost ratio. The results of Sections 3 and 4 are derived for large population size, N. In Section 5 ('Finite population size') we give the analytic expressions for any N. In Section 6 ('Numerical Simulations') we compare our analytical results with computer simulations. In Section 7 ('General Payoff Matrix') we study the competition between two strategies, A and B, for a general payoff matrix. In the Appendix we give the analytic proof of our results.

1 Evolution of cooperation on sets

Consider a population of N individuals distributed over M sets. Each individual belongs to exactly K sets, where $K \leq M$. Additionally, each individual has a strategy $s_i \in \{0, 1\}$, refered to as cooperation, 1, or defection, 0.

The system evolves according to a Wright-Fisher process [1]-[3]. There are discrete, non-verlapping generations. All individuals update at the same time. The population size is constant. Individuals reproduce proportional to their fitness [4], [5]. An offspring inherits the sets of the

parent with probability $1-v$ or adopts a random configuration (including that of the parent) with probability v. Any particular configuration of set memberships is chosen with probability $v/\binom{M}{K}$. Similarly, the offspring inherits the strategy of the parent with probability $1-u$; with probability u he adopts a random strategy. Thus, we have a strategy mutation rate, u, and a set mutation rate, v.

If two individuals belong to the same set, they interact; if they have more than one set in common, they interact several times. An interaction can be any evolutionary game, but first we consider a simplified Prisoner's Dilemma game given by the payoff matrix:

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array} \tag{1}$$

Here $b > 0$ is the benefit gained from cooperators and $c > 0$ is the cost cooperators have to pay. The payoff gained by an individual in each interaction is added to his total payoff, p. The fitness of an individual is $f = 1 + \delta p$, where δ corresponds to the intensity of selection [6]. The limit of weak selection is given by $\delta \to 0$. Neutral drift corresponds to $\delta = 0$.

A state S of the system is given by a vector s and a matrix H. s is the strategy vector; its entry s_i describes the strategy of individual i. Thus, $s_i = 0$ if i is a defector and it is 1 if i is a cooperator. H is an $N \times M$ matrix whose ij-th entry is 1 if individual i belongs to set j and is 0 otherwise. We will refer to row i of H as the vector h_i; this vector gives the set memberships of individual i.

Considering that two individuals interact as many times as many sets they have in common and that we do not allow self-interaction, we can now write the fitness of individual i as

$$f_i = 1 + \delta \sum_{j \neq i} (h_i \cdot h_j)(-cs_i + bs_j) \tag{2}$$

Thus, individual i interacts with individual j only if they have at least one set in common ($h_i \cdot h_j \neq 0$). In this case, they interact as many times as they find themselves in joint sets, which is given by the dot product of their set membership vectors. For each interaction, i pays a cost if he is a cooperator ($s_i = 1$) and receives a benefit if j is a cooperator ($s_j = 1$).

Let F_C be the total payoff for cooperators, $F_C = \sum_i s_i f_i$. Let F_D be the total payoff for defectors, $F_D = \sum_i (1 - s_i) f_i$. The total number of cooperators is $\sum_l s_l$. The total number of defectors is $N - \sum_l s_l$. Provided that the number of cooperators and that of defectors are both non-zero, we can write the average payoff of cooperators and defectors as

$$f_C = \frac{\sum_i s_i f_i}{\sum_l s_l} \qquad\qquad f_D = \frac{\sum_i (1 - s_i) f_i}{N - \sum_l s_l} \tag{3}$$

The average fitness of cooperators is greater than that of defectors, $f_C > f_D$, if

$$\sum_i s_i f_i (N - \sum_l s_l) > \sum_l s_l \sum_i (1 - s_i) f_i \iff \tag{4}$$

$$N \sum_i s_i f_i > \sum_l s_l \sum_i f_i \tag{5}$$

We rewrite equation (2) in a more convenient form, using the fact that $h_i \cdot h_i = K$ for any i.

$$f_i = 1 + \delta \Big(\sum_j h_i \cdot h_j (-c s_i + b s_j) - K(b - c) s_i \Big) \tag{6}$$

Then inequality (5) leads to

$$b \sum_{i,j} s_i s_j h_i \cdot h_j - c \sum_{i,j} s_i h_i \cdot h_j > \frac{b - c}{N} \sum_{i,j,l} s_i s_l h_i \cdot h_j - \frac{K(b - c)}{N} \sum_{i,l} s_i s_l + \frac{K(b - c)}{N} \sum_i s_i \tag{7}$$

In order for cooperators to be favored, we want this inequality to hold on average, where the average is taken in the stationary state, that is over every possible state S of the system, weigthed by the probability π_S of finding the system in each state.

$$b \Big\langle \sum_{i,j} s_i s_j h_i \cdot h_j \Big\rangle - c \Big\langle \sum_{i,j} s_i h_i \cdot h_j \Big\rangle > \frac{b - c}{N} \Big\langle \sum_{i,j,k} s_i s_k h_i \cdot h_j \Big\rangle - \frac{K(b - c)}{N} \Big\langle \sum_{i,k} s_i s_k \Big\rangle + \frac{K(b - c)}{N} \Big\langle \sum_i s_i \Big\rangle \tag{8}$$

The angular brackets denote this average. Thus, for example,

$$\Big\langle \sum_{i,j} s_i s_j h_i \cdot h_j \Big\rangle = \sum_S \Big(\sum_{i,j} s_i s_j h_i \cdot h_j \Big) \cdot \pi_S \tag{9}$$

So far this has been an intuitive derivation. A more rigorous analytic derivation which explains why we take these averages is presented in the Appendix; it also appears in a different context in [7]. We further show in the Appendix that when we take the limit of weak selection, the above condition (7) is equivalent to the one where we take these averages in the neutral stationary state (see equations (90), (92) and the discussion following them). Neutrality means that no game is being played and all individuals have the same fitness. Thus, we show that in the limit of weak selection the decisive condition is

$$b\Big\langle \sum_{i,j} s_i s_j h_i \cdot h_j \Big\rangle_0 - c\Big\langle \sum_{i,j} s_i h_i \cdot h_j \Big\rangle_0 > \frac{b-c}{N}\Big\langle \sum_{i,j,k} s_i s_k h_i \cdot h_j \Big\rangle_0 - \frac{K(b-c)}{N}\Big\langle \sum_{i,k} s_i s_k \Big\rangle_0 + \frac{K(b-c)}{N}\Big\langle \sum_{i} s_i \Big\rangle_0 \tag{10}$$

The zero subscript refers to taking the average in the neutral state, $\delta = 0$. For example, the term $\langle \sum_{i,j} s_i s_j h_i \cdot h_j \rangle_0$ is

$$\Big\langle \sum_{i,j} s_i s_j h_i \cdot h_j \Big\rangle_0 = \sum_S \Big(\sum_{i,j} s_i s_j h_i \cdot h_j \Big) \cdot \pi_S^{(0)} \tag{11}$$

Here $\pi_S^{(0)}$ is the neutral stationary probability that the system is in state S. As before, the sum is taken over all possible states S.

Solving for b/c we obtain the critical benefit-to-cost ratio

$$\left(\frac{b}{c}\right)^* = \frac{\langle \sum_{i,j} s_i h_i \cdot h_j \rangle_0 - \frac{1}{N}\langle \sum_{i,j,l} s_i s_j h_j \cdot h_l \rangle_0 - K\langle \sum_i s_i \rangle_0 + \frac{K}{N}\langle \sum_{i,j} s_i s_j \rangle_0}{\langle \sum_{i,j} s_i s_j h_i \cdot h_j \rangle_0 - \frac{1}{N}\langle \sum_{i,j,l} s_i s_j h_j \cdot h_l \rangle_0 - K\langle \sum_i s_i \rangle_0 + \frac{K}{N}\langle \sum_{i,j} s_i s_j \rangle_0} \tag{12}$$

Thus, we have expressed the critical b/c ratio in the limit of weak selection, only in terms of correlations from the neutral stationary state. Now we can focus on the neutral case to obtain the desired terms. Nevertheless, the results we derive hold in the limit of weak selection, $\delta \to 0$.

The strategies and the set memberships of the individuals change independently. All correlations in (12) can be expressed as averages and probabilities in the stationary state. We will describe each necessary term separately.

First we consider the term

$$\Big\langle \sum_i s_i \Big\rangle_0 = N\Pr(s_i = 1) = \frac{N}{2} \tag{13}$$

This is simply the average number of cooperators. In the absence of selection this is $N/2$.

Next we consider

$$\Big\langle \sum_{i,j} s_i s_j \Big\rangle_0 = N^2 \Pr(s_i = s_j = 1) = \frac{N^2}{2} \Pr(s_i = s_j) \tag{14}$$

The first equality is self-explanatory. The second equality follows from the fact that in the neutral stationary state the two strategies are equivalent. Thus we can interchange any 0 with any 1 and we can express everything in terms of individuals having the same strategy rather than being both cooperators. Thus in the absence of selection, $\Pr(s_i = s_j = 1) = \Pr(s_i = s_j)/2$. We will use the same idea for all terms below.

Next, we consider the term

$$\Big\langle \sum_{i,j} s_i h_i \cdot h_j \Big\rangle_0 = N^2 \langle h_i \cdot h_j \ \mathbf{1}_{s_i=1} \rangle_0 = \frac{N^2}{2} \langle h_i \cdot h_j \rangle_0 \tag{15}$$

The function $\mathbf{1}_{(\cdot)}$ is the indicator function. It is 1 if its argument is true and 0 if it is false. For the last two terms of the equality, i and j are any two individuals picked at random, with replacement.

In the sum $\sum_{i,j} s_i h_i \cdot h_j$, we add the term $h_i \cdot h_j$ only if $s_i = 1$; otherwise, we add 0. Nevertheless, our sum has N^2 terms. This leads to the first equality. To be more precise, we should say that $\langle \sum_{i,j} s_i h_i \cdot h_j \rangle_0 = N^2 \cdot \mathrm{E}[\langle h_i \cdot h_j \ \mathbf{1}_{s_i=1} \rangle_0]$, where the expectation is taken over the possible pairs (i, j). For simplicity, we will omit the expectation symbol.

We can think of the term $\langle h_i \cdot h_j \ \mathbf{1}_{s_i=1} \rangle_0$ as the average number of sets two random individuals have in common given that they have a non-zero contribution to the average only if the first one is a cooperator. By the same reasoning as above, any i can be interchanged with any j and thus we can obtain the second equality in (15). Thus, the term we end up calculating is $\langle h_i \cdot h_j \rangle_0$ which is the average number of sets two random individuals have in common.

The same reasoning leads to the final two correlations. We have

$$\Big\langle \sum_{i,j} s_i s_j h_i \cdot h_j \Big\rangle_0 = N^2 \langle h_i \cdot h_j \ \mathbf{1}_{s_i=s_j=1} \rangle_0 = \frac{N^2}{2} \langle h_i \cdot h_j \ \mathbf{1}_{s_i=s_j} \rangle_0 \tag{16}$$

Using the same wording as above, equation (16) is the average number of sets two individuals have in common given that only individuals with the same strategy have a non-zero contribution

to the average.

Finally, we can write

$$\Big\langle \sum_{i,j,l} s_i s_j h_j \cdot h_l \Big\rangle_0 = N^3 \langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j=1} \rangle_0 = \frac{N^3}{2} \langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \rangle_0 \tag{17}$$

For this term we need to pick three individuals at random, with replacement. Then, (17) is the average number of sets the latter two have in common, given that they have a non-zero contribution to the average only if the first two are cooperators.

Therefore, we can rewrite the critical ratio as

$$\left(\frac{b}{c}\right)^* = \frac{N\langle h_i \cdot h_j \rangle_0 - N\langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \rangle_0 - K + K \cdot \Pr(s_i = s_j)}{N\langle h_i \cdot h_j \ \mathbf{1}_{s_i=s_j} \rangle_0 - N\langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \rangle_0 - K + K \cdot \Pr(s_i = s_j)} \tag{18}$$

For simplicity, we want to find the above quantities when the three individuals are chosen without replacement. We know, however, that out of two individuals, we pick the same individual twice with probability $1/N$. Moreover, given three individuals i, j and l, the probability that two are the same but the third is different is $(1/N)(1-1/N)$, whereas the probability that all three are identical is $1/N^2$.

Let us make the following notation

$$y = \Pr(s_i = s_j \mid i \neq j) \tag{19}$$

$$z = \langle h_i \cdot h_j \mid i \neq j \rangle_0 \tag{20}$$

$$g = \langle h_i \cdot h_j \ \mathbf{1}_{s_i=s_j} \mid i \neq j \rangle_0 \tag{21}$$

$$h = \langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \mid i \neq j \neq l \rangle_0 \tag{22}$$

Then the quantities of interest in (18) become

$$\Pr(s_i = s_j) = \frac{1}{N}\frac{N}{2} + \frac{N-1}{N} y \tag{23}$$

$$\langle h_i \cdot h_j \rangle_0 = K\frac{1}{N} + \frac{N-1}{N} z \tag{24}$$

$$\langle h_i \cdot h_j \ \mathbf{1}_{s_i=s_j} \rangle_0 = K\frac{1}{N} + \frac{N-1}{N} g \tag{25}$$

$$\langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \rangle_0 = K\frac{1}{N^2} + \frac{(N-1)(N-2)}{N^2} h + \frac{N-1}{N^2}(z + g + Ky) \tag{26}$$

The critical ratio can now be expressed in terms of z, g and h as

$$\left(\frac{b}{c}\right)^* = \frac{(N-2)(z-h)+z-g}{(N-2)(g-h)-z+g} \tag{27}$$

Note that y cancels. In the limit $N \to \infty$ we have

$$\left(\frac{b}{c}\right)^* = \frac{z-h}{g-h} \tag{28}$$

For calculating the critical benefit-to-cost ratio in the limit of weak selection, it suffices to find z, g and h in the neutral case: z is the average number of sets two randomly picked individuals have in common; g is the average number of sets they have in common given that only individuals with the same strategy have a non-zero contribution to the average. For h we need to pick three individuals at random; then h is the average number of sets the latter two have in common given that they have a non-zero contribution to the average only if the first two have the same strategy.

In general these quantities cannot be written as independent products of the average number of common sets times the probability of having the same strategy. However, if we fix the time to their most recent common ancestor (MRCA), then the set mutations and strategy mutations are obviously independent. If we knew that the time to their MRCA is $T = t$, we could then write g, for instance, as the product:

$$\langle h_i \cdot h_j \mid i \neq j,\ T = t \rangle_0 \cdot \Pr(s_i = s_j \mid i \neq j,\ T = t) \tag{29}$$

Two individuals always have a common ancestor if we go back in time far enough. However, we cannot know how far we need to go back. Thus, we have to account for the possibility that $T = t$ takes values anywhere between 1 and ∞. Note that $T = 0$ is excluded because we assume that the two individuals are distinct. Moreover, we know that this time is affected neither by the strategies, nor by the set memberships of the two individuals. It is solely a consequence of the W-F dynamics.

We can calculate z, g and h provided that we know that the time to their MRCA is T. We first calculate the probability that given two random individuals, i and j, their MRCA is at time $T = t$:

$$\Pr(T = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \tag{30}$$

Next, we calculate the probability that given three randomly picked individuals i, j and k, and looking back at their trajectories, the first merging happens at time t_3 while the second one takes t_2 more time steps. If we follow the trajectories of these individuals back in time, the probability that there was no coalescence event in one time step is $(1-1/N)(1-2/N)$. Two individuals coalesce with probability $3/N(1-1/N)$. When two individuals have coalesced, the remaining two merge with probability $1/N$ during an update step. For the probability that the first merging happens at time $t_3 \geq 1$ and the second takes $t_2 \geq 1$ more time steps, we obtain

$$\Pr(t_3, t_2) = \frac{3}{N^2}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\right]^{t_3-1}\left(1-\frac{1}{N}\right)^{t_2} \tag{31}$$

The probability that all three paths merge simultaneously at time t_3 is:

$$\Pr(t_3, 0) = \frac{1}{N^2}\left[\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\right]^{t_3-1} \tag{32}$$

We can calculate both the case of finite N and the limit $N \to \infty$. The results for finite N will be discussed in Section 5. Here we deal only with the $N \to \infty$ limit. In this case, we introduce the notations $\tau = t/N$, $\tau_2 = t_2/N$ and $\tau_3 = t_3/N$. We use a continuous time description, with τ, τ_2, τ_3 ranging between 0 and ∞. In the continuous time limit, the coalescent time distributions in (30) and (31) are given by

$$p(\tau) = e^{-\tau} \tag{33}$$

and

$$p(\tau_3, \tau_2) = 3e^{-(3\tau_3+\tau_2)} \tag{34}$$

Due to the non-overlapping generations of the W-F process, each individual is newborn and has the chance to mutate both in strategy and in set configuration. In the $N \to \infty$ and $u, v \to 0$ limits, we can consider our process to be continuous in time, with strategy mutations arriving at a rate $\mu = 2Nu$ and set membership mutations arriving at a rate $\nu = 2Nv$ in the ancestry line of any two individuals.

2 Belonging to K sets

First, we find the probability that two individuals have the same strategy at time t from their MRCA. Next we find the average number of sets two individuals have in common, a quantity which is necessary for finding z, g and h. Finally, we calculate the critical benefit-to-cost ratio.

2.1 Probability that two individuals have the same strategy

The first quantity we need is the probability that two individuals have the same strategy at time t from their MRCA. Imagining the two paths of two individuals i and j starting from their MRCA, it is easy to note that two players have the same strategy at time t if the total number of mutations which occurred in their ancestry lines is even. The probability that an even number of mutations has occurred is

$$y(t) = \Pr(s_i = s_j \mid T = t) = \sum_{l=0}^{t} \binom{2t}{2l} \left(1 - \frac{u}{2}\right)^{2t-2l} \left(\frac{u}{2}\right)^{2l} = \frac{1 + (1-u)^{2t}}{2} \tag{35}$$

In the continuous time limit, making the substitutions $\tau = t/N$ and $\mu = 2Nu$, we obtain

$$y(\tau) = \lim_{N\to\infty} \frac{1 + \left(1 - \frac{\mu}{2N}\right)^{2\tau N}}{2} = \frac{1 + e^{-\mu\tau}}{2} \tag{36}$$

The limits above are taken for $\mu =$ constant.

2.2 Average number of sets two individuals have in common

The first quantity we need is $z = \langle h_i \cdot h_j \mid i \neq j \rangle_0$. As in the previous subsection, we begin by calculating this probability given that the MRCA of the two individuals is at time τ. We use the notation

$$z(\tau) = \langle h_i \cdot h_j \mid i \neq j,\ T = \tau \rangle_0 \tag{37}$$

We can interpret z as the average number of sets two randomly chosen, distinct individuals have in common. Then $z(\tau)$ is this same average, but taken now only over the states where $T = \tau$.

We start by finding the probability that two such individuals have $0 \leq i \leq K$ sets in common. We then have two options at time τ from their MRCA: neither of the two have changed their

configuration or at least one of them has changed his configuration. In the second case, the two individuals become random in terms of the comparison of their set memberships.

Thus, we analyze the following possibilities:

- neither has changed with probability $e^{-\nu\tau}$ and in this case they still have K sets in common;
- at least one has changed with probability $1-e^{-\nu\tau}$ and in this case they can have $i \in \{0, \dots, K\}$ sets in common with probability:

$$\binom{K}{i} \frac{\binom{M-K}{K-i}}{\binom{M}{K}} \tag{38}$$

Hence, the probability that two individuals have $i \leq K$ sets in common at time $T = \tau$ from their MRCA is

$$\pi_i(\tau) = \begin{cases} e^{-\nu\tau} + (1 - e^{-\nu\tau})/\binom{M}{K} & \text{if } i = K \\ (1 - e^{-\nu\tau})\binom{K}{i}\binom{M-K}{K-i}/\binom{M}{K} & \text{if } i < K \end{cases} \tag{39}$$

Our goal is to calculate the average number of sets they have in common. We obtain

$$\begin{aligned} z(\tau) &= \sum_{i=1}^{K} i \cdot \pi_i(\tau) = K e^{-\nu\tau} + (1 - e^{-\nu\tau}) \sum_{i=1}^{K-1} i \binom{K}{i} \frac{\binom{M-K}{K-i}}{\binom{M}{K}} \\ &= e^{-\nu\tau} \left(K - \frac{K^2}{M} \right) + \frac{K^2}{M} \end{aligned} \tag{40}$$

2.3 Finding the critical ratio

Once we know the average number of sets two individuals have in common at time τ from their MRCA, we can again use the method of the coalescent to express $z = \langle h_i \cdot h_j \mid i \neq j \rangle_0$ in the continuous time limit as

$$z = \int_0^\infty z(\tau) p(\tau) \, \mathrm{d}\tau = \frac{K(M + \nu K)}{M(\nu + 1)} \tag{41}$$

The next quantity we need is $g = \langle h_i \cdot h_j \, \mathbf{1}_{s_i = s_j} \mid i \neq j \rangle_0$. As explained above, this can be interpreted as the average number of sets two distinct random individuals have in common given that they have a non-zero contribution to the average only if they have the same strategy. Let $g(\tau) = \langle h_i \cdot h_j \, \mathbf{1}_{s_i = s_j} \mid i \neq j, \, T = \tau \rangle_0$. Once we fix the MRCA, the set mutations and strategy mutations are independent. Thus $g(\tau)$ can be written as a product of the probability that i and j have the same strategy and the average number of sets two distinct individuals have in common.

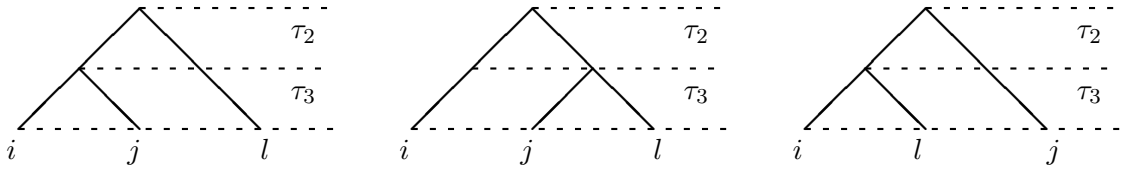
We can then write $g(\tau)$ as

$$g(\tau) = z(\tau)y(\tau) \tag{42}$$

In the continuous time limit, we have

$$g = \int_0^\infty z(\tau)y(\tau)p(\tau)\ \mathrm{d}\tau = \frac{K}{2M}\left(\frac{M+\nu K}{1+\nu} + \frac{K}{1+\mu} + \frac{M-K}{1+\nu+\mu}\right) \tag{43}$$

Finally, we need to find $h = \langle h_j \cdot h_l\ \mathbf{1}_{s_i=s_j} \mid i \neq j \neq l\rangle_0$. This can be interpreted as follows: we pick three distinct random individuals i, j and l and ask how many sets j and l have in common on average, given that they have a non-zero contribution to the average only if i and j have the same strategy. As before, we need to fix the time up to the MRCA of the particular pairs of individuals. Let $T(i,j)$ (respectively $T(j,l)$) be the time up to the MRCA of i and j (respectively j and l). If we look back far enough, all three individuals will have an MRCA (all their paths will coalesce). However, we do not know which two paths coalesce first, so we have to analyze all possibilities (shown in the figure below). Let τ_3 be the time up to the first coalescence and let τ_2 be the time between the first and the second coalescence. Then, we can find $T(i,j)$ and $T(j,l)$ in each case



i and j coalesce first	j and l coalesce first	i and l coalesce first
$T(i,j) = \tau_3$	$T(i,j) = \tau_2 + \tau_3$	$T(i,j) = \tau_2 + \tau_3$
$T(j,l) = \tau_2 + \tau_3$	$T(j,l) = \tau_3$	$T(j,l) = \tau_2 + \tau_3$

The possibility that all three coalesce simultaneously is included in these three cases (when $\tau_2 = 0$).

Let $h(\tau_3, \tau_2) = \langle h_j \cdot h_l\ \mathbf{1}_{s_i=s_j} \mid i \neq j \neq l\rangle_0$. Once we fix the times to the MRCA's, the set mutations and the strategy mutations become independent. Then we can write $h(\tau_3, \tau_2)$ as a product. We already know the probability $y(\tau)$ that two individuals have the same strategy at time τ from their MRCA (36). Moreover, we know $z(\tau)$, the average number of sets they have in common if the time to their MRCA is $T = \tau$. So we only need to use the times $T(i,j)$ and $T(j,l)$ calculated in each of the three cases.

With probability 1/3 we are in the first case, where i and j coalesce first and then they coalesce with l. In this case, we can write $h(\tau_3, \tau_2) = y(\tau_3)z(\tau_3 + \tau_2)$. With probability 1/3 we are in the second case, where j and l coalesce first and then they coalesce with i. Then $h(\tau_3, \tau_2) = y(\tau_3 + \tau_2)z(\tau_3)$. Finally, with probability 1/3 we are in the last case, where i and l coalesce first and then they coalesce with j. In this case $h(\tau_3, \tau_2) = y(\tau_3 + \tau_2)z(\tau_3 + \tau_2)$.

We finally obtain the expression for $h = \langle h_j \cdot h_l \ \mathbf{1}_{s_i=s_j} \mid i \neq j \neq l \rangle_0$ by adding the values in all three cases, multiplied by the corresponding probabilities.

$$\begin{aligned} h &= \frac{1}{3}\int_0^\infty \mathrm{d}\tau_3 \int_0^\infty \mathrm{d}\tau_2 \ [y(\tau_3)z(\tau_3+\tau_2) + y(\tau_3+\tau_2)z(\tau_3) + y(\tau_3+\tau_2)z(\tau_3+\tau_2)]p(\tau_3,\tau_2) \\ &= \frac{\nu K^2(3+10\mu+6\mu^2+\mu^3+\nu^2(2+\mu)+2\nu(2+\mu)^2)}{2M(1+\nu)(1+\mu)(1+\nu+\mu)(3+\nu+\mu)} + \\ &+ \frac{MK(3+11\mu+6\mu^2+\mu^3+\nu^2(2+\mu)+\nu(8+9\mu+2\mu^2))}{2M(1+\nu)(1+\mu)(1+\nu+\mu)(3+\nu+\mu)} \end{aligned} \tag{44}$$

Now we have calculated d, g and h. We can use (28) to obtain the critical b/c ratio

$$\left(\frac{b}{c}\right)^* = \frac{K}{M-K}(\nu+2+\mu) + \frac{M}{M-K}\frac{\nu^2+3\nu+3+2(\nu+2)\mu+\mu^2}{\nu(\nu+2+\mu)} \tag{45}$$

Note that the critical b/c ratio depends only on M/K and not on both M and K.

For $\mu \to 0$ we have

$$\left(\frac{b}{c}\right)^* = \frac{K}{M-K}(\nu+2) + \frac{M}{M-K}\frac{\nu^2+3\nu+3}{\nu(\nu+2)} \tag{46}$$

If the benefit-to-cost ratio exceeds this value, then cooperators are more frequent than defectors in the equilibrium distribution of the mutation-selection process.

3 Triggering cooperation

We will now study an extended model, where cooperation is only triggered if the other person has L sets in common. We have $1 \leq L \leq K$. Setting $L = 1$ takes us back to the previous framework.

In order to account for this conditional interaction, we define the following variable:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } h_i \cdot h_j \geq L \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

The fitness of individual i can be written as

$$f_i = 1 + \delta \sum_{j \neq i} \gamma_{ij} h_i \cdot h_j (-cs_i + bs_j) \tag{48}$$

Everything follows exactly as before, with the only change that wherever we have the product $h_i \cdot h_j$, it will be replaced by $\gamma_{ij} h_i \cdot h_j$. The quantity $y(\tau)$ remains unchanged and represents the probability that two random, distinct individuals have the same strategy at time τ from their MRCA. The quantity that is affected by the change is $z(\tau)$ which now becomes $z(\tau) = \langle \gamma_{ij} h_i \cdot h_j \rangle_0$; this is now the average number of sets two individuals have in common, provided that they have at least L sets in common. Implicitly, z, g and h will change: instead of accounting for the average number of sets in common, they account for the average number of sets in common, provided that there are at least L common sets.

As before, we can express z, g and h as follows

$$z = \int_0^\infty z(\tau) p(\tau) \, \mathrm{d}\tau \tag{49}$$

$$g = \int_0^\infty \Pr(s_i = s_j \mid T = \tau) \langle \gamma_{ij} h_i \cdot h_j \mid T = \tau \rangle_0 p(\tau) \, \mathrm{d}\tau = \int_0^\infty z(\tau) y(\tau) p(\tau) \, \mathrm{d}\tau \tag{50}$$

$$h = \frac{1}{3} \int_0^\infty \mathrm{d}\tau_3 \int_0^\infty \mathrm{d}\tau_2 \, [y(\tau_3) z(\tau_3 + \tau_2) + y(\tau_3 + \tau_2) z(\tau_3) + y(\tau_3 + \tau_2) z(\tau_3 + \tau_2)] p(\tau_3, \tau_2) \tag{51}$$

We obtain the same expressions (27) and (28) in terms of these new z, g and h.

The only quantity that has changed and needs to be calculated is $z(\tau) = \langle \gamma_{ij} h_i \cdot h_j \mid T = \tau \rangle_0$. The reasoning is as before, but we now need to account for the γ_{ij}. We start by finding the probability that two random individuals have $0 \leq i \leq K$ sets in common. This follows exactly as before: the probability that two individuals have $i \leq K$ sets in common at time $T = \tau$ from their MRCA is

$$\pi_i(\tau) = \begin{cases} e^{-\nu\tau} + (1 - e^{-\nu\tau}) / \binom{M}{K} & \text{if } i = K \\ (1 - e^{-\nu\tau}) \binom{K}{i} \binom{M-K}{K-i} / \binom{M}{K} & \text{if } i < K \end{cases} \tag{52}$$

Our goal is to estimate the quantity $z(\tau) = \langle \gamma_{ij} h_i \cdot h_j \mid T = \tau \rangle_0$. We know that $\gamma_{ij} = 0$ if they have less than L sets in common. Only the cases when they have at least L sets in common contribute to our average. Therefore, we have

$$z(\tau) = \langle \gamma_{ij} h_i \cdot h_j \mid T = \tau \rangle_0 = \sum_{i=L}^{K} i \cdot \pi_i(t) \tag{53}$$

We have already analyzed the case $L = 1$. We will now study the case $1 < L \leq K$. We denote:

$$K^* = \frac{M}{K} \sum_{i=L}^{K} i \binom{K}{i} \binom{M-K}{K-i} / \binom{M}{K} = K \sum_{i=L}^{K} \binom{K-1}{i-1} \binom{M-K}{K-i} / \binom{M-1}{K-1} \tag{54}$$

Note that K^* need not be an integer and that for $L = 1$ we obtain $K^* = K$.

Using (53) and (39), we can rewrite $z(\tau)$ as

$$z(\tau) = K(1 - \frac{K^*}{M}) e^{-\nu\tau} + K \frac{K^*}{M} \tag{55}$$

The critical ratio (28) becomes

$$\left(\frac{b}{c}\right)^* = \frac{K^*}{M - K^*}(\nu + 2 + \mu) + \frac{M}{M - K^*} \frac{\nu^2 + 3\nu + 3 + 2(\nu + 2)\mu + \mu^2}{\nu(\nu + 2 + \mu)} \tag{56}$$

Since for $L = 1$ we have $K^* = K$, we recover the previous result, (45). For $\mu \to 0$ we have

$$\left(\frac{b}{c}\right)^* = \frac{K^*}{M - K^*}(\nu + 2) + \frac{M}{M - K^*} \frac{\nu^2 + 3\nu + 3}{\nu(\nu + 2)} \tag{57}$$

The critical benefit-to-cost ratio is the same as given by equation (46) but now K is replaced by K^*, which is the 'effective' number of sets that each individual belongs to. The smaller K^* is, the smaller is the critical benefit-to-cost ratio. The critical benefit-to-cost ratio depends on M/K^*.

The mechanism for the evolution of cooperation in our model is similar to the clustering that occurs in spatial games [8] and games on graphs [9], [10]. Cooperators cluster together in some sets and thereby gain an advantage over defectors. But the difference between evolutionary graph theory and evolutionary set theory is the following. In evolutionary graph theory, both the interaction (which leads to payoff) and the evolutionary updating must be local for cooperators to evolve. In evolutionary set theory, the interactions are local (among members of the same set), but the

evolutionary updating is global: every individual can choose to imitate every other individual.

Our model is also very different from standard models of group selection. In standard models of group selection, (i) each individual belongs to one group; (ii) there is selection on two levels (that of the individual and that of the group); (iii) and there is competition between groups resulting in turnover of groups. In contrast, in evolutionary set theory, (i) individuals can belong to several sets; (ii) there is only selection on the level of the individual; (iii) and there is no turnover of sets.

4 The minimum benefit-to-cost ratio

The critical b/c ratio given by eqn (57) has a minimum as a function of ν. To find this minimum, we differentiate $(b/c)^*$ as a function of ν and set the result equal to zero. We obtain

$$\frac{M}{K^*} = \nu^2 \cdot \frac{\nu^2 + 4\nu + 4}{\nu^2 + 6\nu + 6} \tag{58}$$

It is easy to show that the solution, ν_{opt}, of this equation satisfies the inequality $\sqrt{\frac{M}{K^*}} < \nu_{opt} < \sqrt{\frac{M}{K^*}} + 1$. Consequently, when M/K^* is large, the optimum ν is

$$\nu_{opt} = \sqrt{\frac{M}{K^*}} \tag{59}$$

Thus, for M/K^* large, we obtain

$$\left(\frac{b}{c}\right)_{min} = 1 + 2\sqrt{\frac{K^*}{M}} \tag{60}$$

Figure S1 shows the critical benefit-to-cost ratio as a function of the effective set mutation rate $\nu = 2Nv$ for various choices of K and L. We use $M = 10$, $N \to \infty$ and $u \to 0$. As a function of ν, the benefit-to-cost ratio is a U-shaped function. If the set mutation rate is too small, then all individuals belong to the same sets. If the set mutation rate is too large, then set affiliations do not persist long enough in time. In both cases the population behaves as if it were well-mixed and hence cooperators have difficulties to thrive. In between, there is an optimum set mutation rate given by (59).

5 Finite population size

In order to do the exact calculation for finite population size, we start from equation (27):

$$\left(\frac{b}{c}\right)^* = \frac{(N-2)(z-h)+z-g}{(N-2)(g-h)-z+g}$$

We deal from the beginning with the general case, $1 \leq L \leq K$. First we calculate the probability that two individuals have the same strategy at time t from their MRCA. This is given in (35) as

$$y(t) = \frac{1+(1-u)^{2t}}{2} \tag{61}$$

Next we calculate $z(t) = \langle \gamma_{ij} h_i \cdot h_j \mid T = t \rangle_0$. The reasoning is the same as in the continuous time case. We analyze the following possibilities:

- neither has changed with probability $(1-v)^{2t}$ and in this case they still have K sets in common;
- at least one has changed with probability $1-(1-v)^{2t}$ and in this case they can have $i \in \{0, \ldots, K\}$ sets in common with probability

$$\binom{K}{i} \frac{\binom{M-K}{K-i}}{\binom{M}{K}} \tag{62}$$

Hence, the probability that two individuals have $i \leq K$ sets in common at time $T = t$ from their MRCA is given by

$$\pi_i(t) = \begin{cases} (1-v)^{2t} + (1-(1-v)^{2t})/\binom{M}{K} & \text{if } i = K \\ (1-(1-v)^{2t})\binom{K}{i}\binom{M-K}{K-i}/\binom{M}{K} & \text{if } i < K \end{cases} \tag{63}$$

We obtain:

$$z(t) = \langle \gamma_{ij} h_i \cdot h_j \mid T = t \rangle_0 = \sum_{i=L}^{K} i \cdot \pi_i(t) = (K - K\frac{K^*}{M})(1-v)^{2t} + K\frac{K^*}{M} \tag{64}$$

This gives $z = \langle \gamma_{ij} h_i \cdot h_j \rangle_0$, taking into account the fact that t ranges between 1 and ∞:

$$z = \sum_{t=1}^{\infty} \langle \gamma_{ij} h_i \cdot h_j \mid T = t \rangle_0 Pr(T = t) = \sum_{t=1}^{\infty} d(t) \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \tag{65}$$

We find g and h similarly and use (27) to find the exact critical b/c in the $u \to 0$ limit

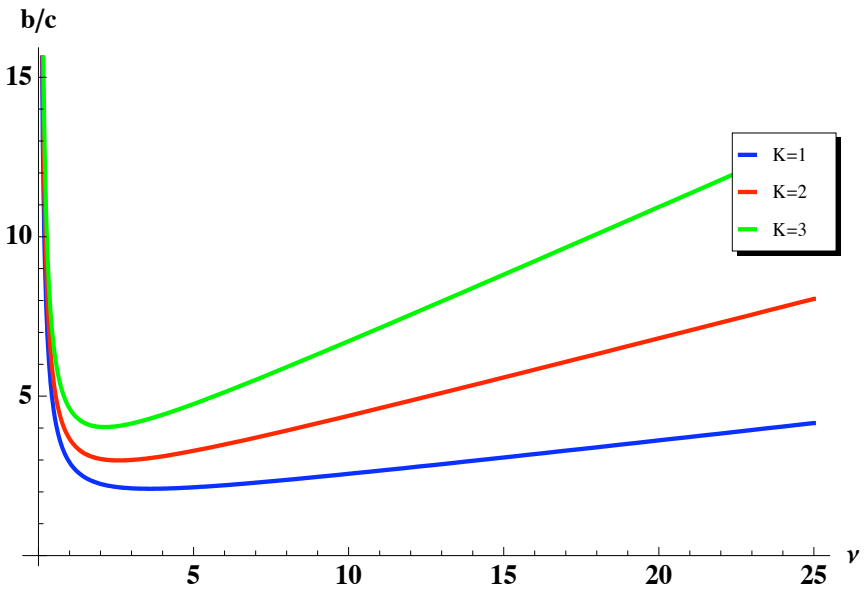
$$\left(\frac{b}{c}\right)^* = \frac{K^* \alpha_1 + M \alpha_2}{K^* \alpha_3 + M \alpha_4} \tag{66}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are the following polynomials in v and N

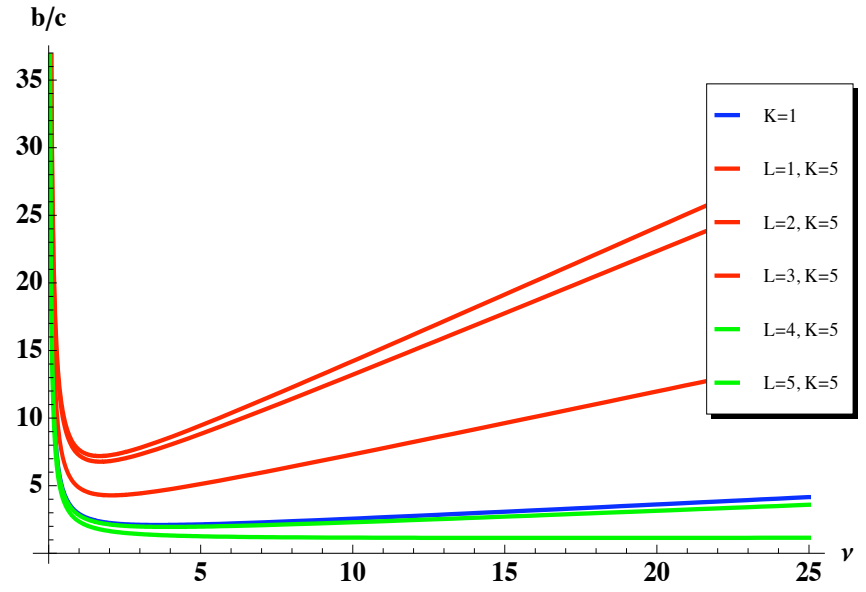
$$\begin{aligned}
\alpha_1 &= (N-1)N^3 v(2-v)[v(2-v)(-Nv(2-v) + 4(1-v)^2) + (1-v)^2(5v^2 - 10v + 4) \\
\alpha_2 &= -(N-1)(1-v)^2[N^2 v(2-v)(-Nv(2-v) - 4v^2 + 8v - 3) - \\
&\quad - (1-v)^2(N(5v^2 - 10v + 3) - 2(1-v)^2)] \\
\alpha_3 &= -v(2-v)[N^4 v(2-v) + N^2(1-v)^2(2N-1)] \\
\alpha_4 &= (1-v)^4(2(1-v)^2 - N(7v^2 - 14v + 3)) - N^2 v(2-v)(1-v)^2(-N^2 v(2-v) - \\
&\quad - N(5v^2 - 10v + 2) + 9v^2 - 18v + 8)
\end{aligned}$$

The exact benefit-to-cost formula (66) gives perfect agreement with our numerical simulations; see Fig. 3 of the main paper.

Figures S2 and S3 illustrate the critical benefit-to-cost ratio as a function of the population size N and of the number of sets M for various choices of K and L. We use $v = 0.01$ and $u = 0.0002$. As a function of N (Fig. S2), the benefit-to-cost ratio is a U-shaped function. If the population size is too small then the effect of spite is too strong. In the limit of $N = 2$, it never pays to cooperate. If the population size is too large (for a fixed number of sets M and a fixed set mutation rate v), then all the sets get populated by defectors and cooperators can not survive. As a function of the number of sets M (Fig. S3), the benefit-to-cost ratio is a declining function. Adding more sets is always helpful for the cooperators.

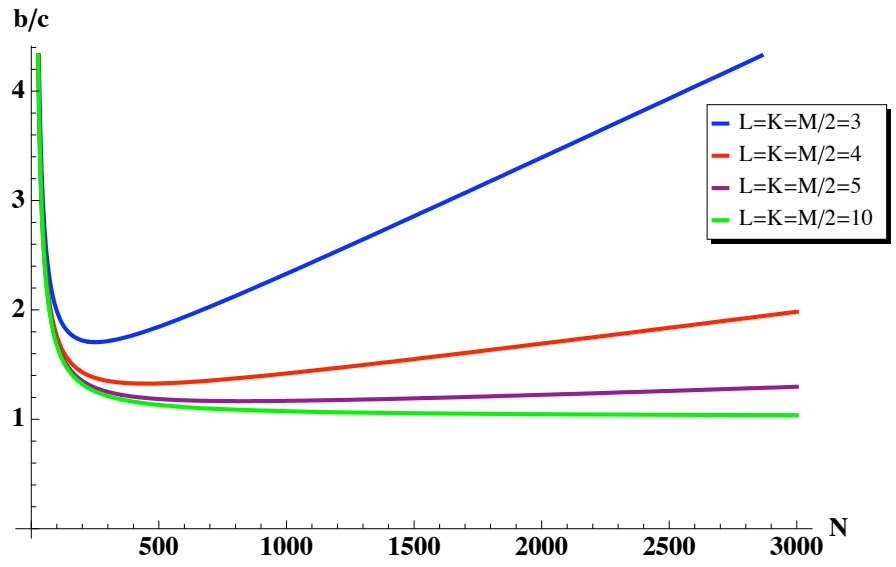


(a)

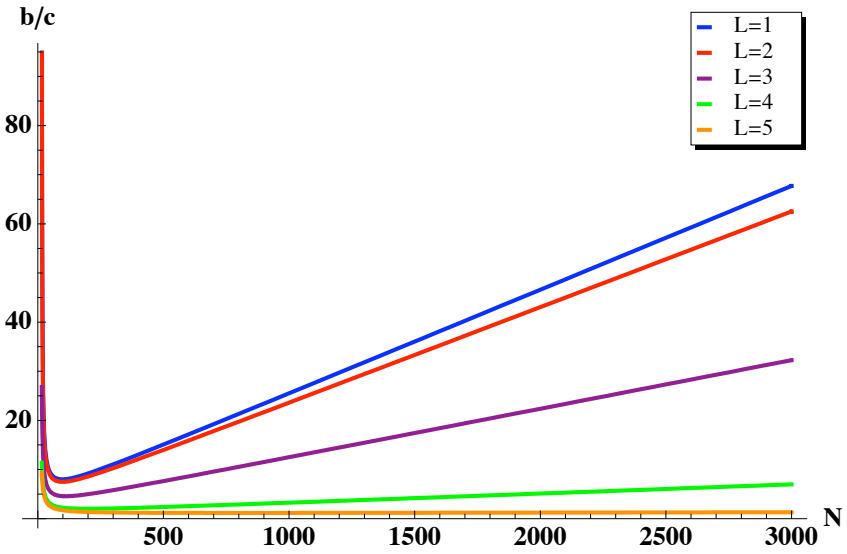


(b)

Fig. S 1: Critical benefit-to-cost ratio as a function of the effective set mutation rate $\nu = 2Nv$. The strategy mutation rate is $u = 0.0002$ and the population size is large, $N \to \infty$. **(a)** Critical b/c ratios for $L = 1, K = 1, 2, 3$, and total number of sets $M = 10$. This shows that increasing K is worse for cooperation. **(b)** Critical b/c ratios for $K = 1$ and for $L = 1, \ldots, 5$, $K = 5$, and $M = 10$. This shows that larger K can be better for cooperation, as long as L is sufficiently large.

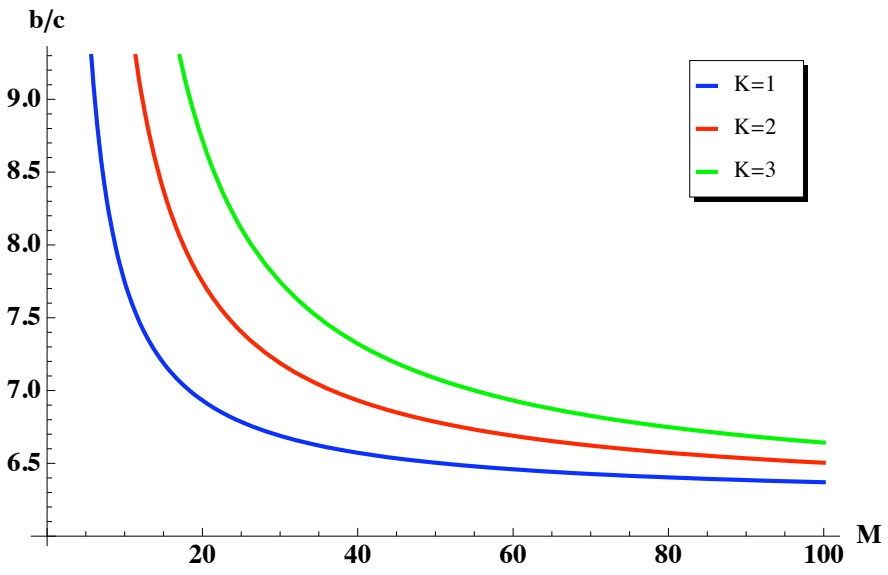


(a)

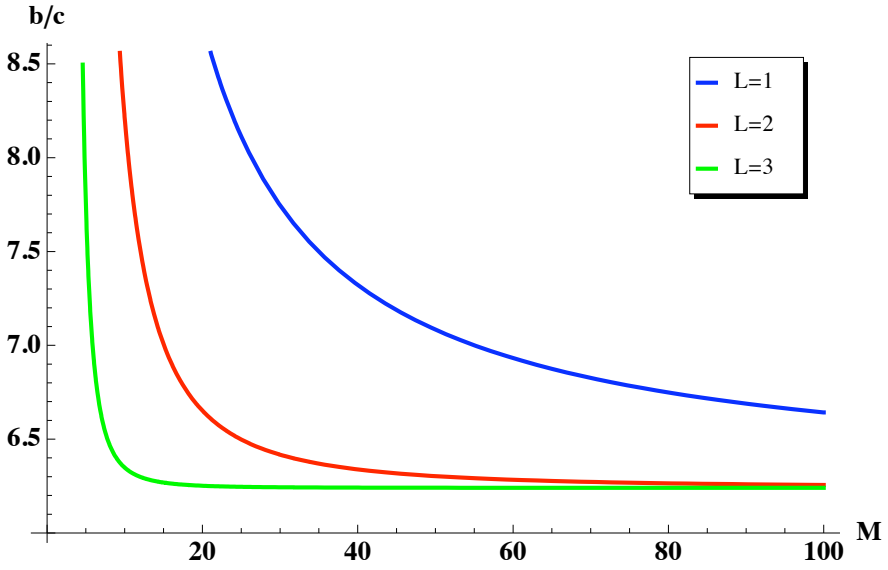


(b)

Fig. S 2: Critical benefit-to-cost ratio as a function of the population size N. The set mutation rate is $v = 0.01$. The strategy mutation rate is $u = 0.0002$. **(a)** Critical b/c ratios for $L = K = M/2$ and $M = 6, 8, 10, 20$. **(b)** Critical b/c ratios for $L = 1, 2, 3, 4, 5, K = 5$ and $M = 10$.



(a)



(b)

Fig. S 3: Critical benefit-to-cost ratio as a function of the number of sets M. The set mutation rate is $v = 0.01$. The strategy mutation rate is $u = 0.0002$. **(a)** Critical b/c ratios for $L = 1, K = 1, 2, 3$ and $N = 20$. **(b)** Critical b/c ratios for $L = 1, 2, 3, K = 3$ and $N = 20$.

6 Numerical simulations

We have performed numerical simulations in order to test the results of our analytical theory (see Figure 3 of the main paper and Figure S3 here). We consider a population of N individuals distributed over M sets. Each individual is in K sets. There are two types of strategies: cooperators, C, and defectors, D. Cooperators pay a cost, c, for another individual to receive a benefit, b. The fitness of an individual is given by $1 + \delta P$, where P is the payoff of the individual. We simulate the Wright-Fisher process for a given intensity of selection, δ.

In each generation we compute the fitness of each individual. The total population size, N, is constant. For the next generation, each individual leaves offspring proportional to its fitness. Thus, selection is always operating. Reproduction is subject to a strategy mutation rate, u, and a set mutation rate v, as explained previously. We follow the population over many generations in order to calculate an accurate time average of the frequency of cooperators in this mutation-selection process.

In Figure S3 we show the time average of the frequency of cooperators as a function of the benefit-to-cost ratio, b/c. We simulate the Wright-Fisher Process for four different intensities of selection ranging from $\delta = 0.05$ to 0.4. The red points indicate the average frequency of cooperators in the mutation selection process. Each point is an average over $t = 5 \times 10^8$ generations. As expected the average frequency of cooperators increases as a function of b/c. We are interested in the value of b/c when cooperators become more abundant than defectors (when their frequency exceeds 1/2.). The vertical blue line is the critical b/c ratio that is predicted from our analytical theory for the limit of weak selection, $\delta \to 0$. We observe excellent agreement between theory and simulations. For weaker intensity of selection the critical value moves closer to the theoretical prediction.

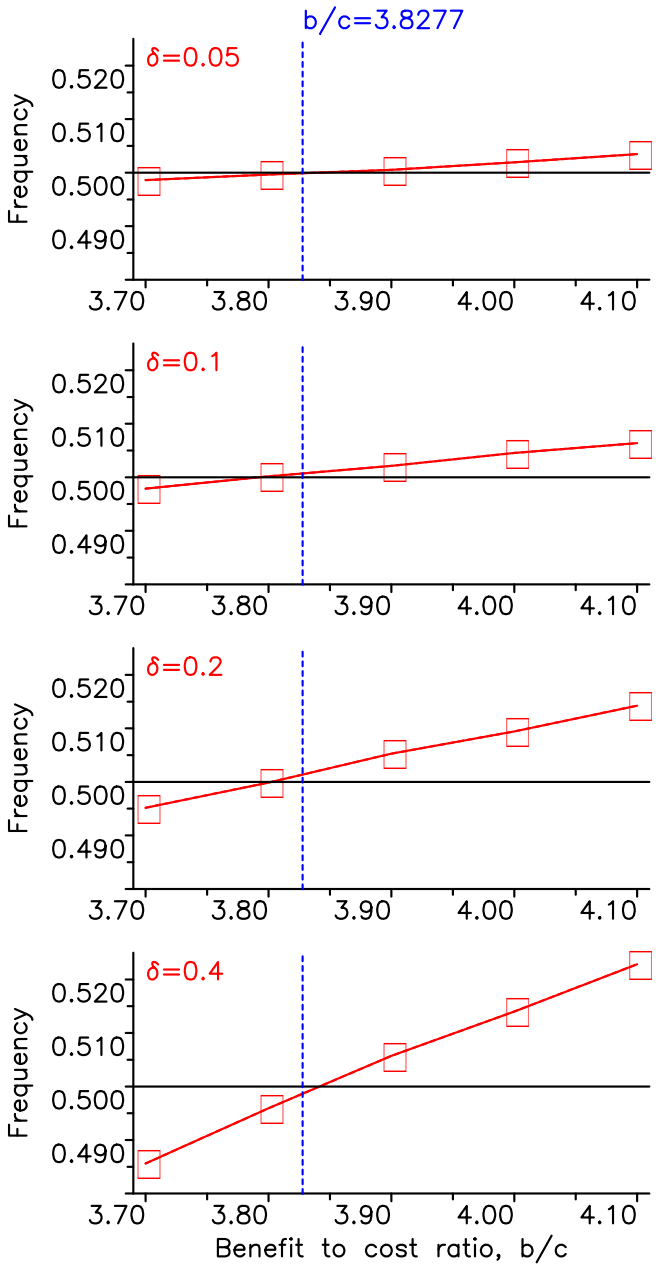


Fig. S 4: Numerical simulation of the mutation-selection process for population size $N = 80$ and various intensities of selection ranging from $\delta = 0.05$ to $\delta = 0.4$. The red dots indicate the frequency of cooperators averaged over $t = 5 \times 10^8$ generations. The cooperator frequency increases as a function of the b/c ratio. For a certain ratio, cooperators become more abundant than defectors (intersection with black horizontal line). The blue vertical line is the theoretical prediction of the critical b/c ratio in the limit of weak selection, $(b/c)^* = 3.8277$. Parameter values: population size $N = 80$, number of sets $M = 10$, number of set memberships $K = 1$, set mutation rate $v = 0.1$, strategy mutation rate $u = 0.004$. We fix $b = 1$ and vary c.

7 General Payoff Matrix

Let us now study a general game between two strategies A and B given by the payoff matrix

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array} \tag{67}$$

A derivation similar to the one presented in the Appendix leads to the following condition necessary for strategy A to be favored over B

$$(R-S)g + (S-P)z > (R-S-T+P)\eta + (S+T-2P)h \tag{68}$$

The quantities z, g and h are as before and η is defined as follows

$$\eta = \langle \gamma_{jl} h_j \cdot h_l \ \mathbf{1}_{s_i=s_j=s_l} \mid i \neq j \neq l \rangle \tag{69}$$

To interpret η, we pick three distinct individuals randomly. Then η is the average number of sets the last two individuals have in common given that they hav a non-zero contribution to the average only if all three have the same strategy. To calculate η we proceed similarly as for h. We fix the times τ_3 up to the first coalescence and τ_2 extra steps up to the second. We denote by $y(\tau_3, \tau_2)$ the probability that the three individuals have the same strategy given these times to the coalescence. We can now rewrite η as

$$\eta = \int_0^\infty \mathrm{d}\tau_2 \int_0^\infty e^{-3\tau_3-\tau_2} y(\tau_3,\tau_2)[z(\tau_3) + z(\tau_2+\tau_3) + z(\tau_2+\tau_3)]\mathrm{d}\tau_3 \tag{70}$$

The only quantity we need to compute is $y(\tau_3, \tau_2)$. Let us assume that i and j coalesce first, and then they coalesce with l. As before, in order for i and j to have the same strategy, the total number of mutations that happen up to their coalescence must be even. Therefore, either both underwent an even number of mutations from their MRCA, or an odd one. If i underwent an odd number, then in order for i and l to have the same number of mutations, it must be that, in the remaining total time (which is $\tau_2 + 2\tau_3$) there must be an odd number of mutations. Thus, in this

case we can write the probability that all three have the same strategy as

$$\frac{1}{8}(1-e^{-\frac{\mu}{2}\tau_3})^2(1-e^{-\frac{\mu}{2}(\tau_3+\tau_2)}) \tag{71}$$

When i undergoes an even number of mutations up to its coalescence with j, the probability that all three have the same strategy is similarly obtained as

$$\frac{1}{8}(1+e^{-\frac{\mu}{2}\tau_3})^2(1+e^{-\frac{\mu}{2}(\tau_3+\tau_2)}) \tag{72}$$

Thus, we can write

$$y(\tau_3,\tau_2)=\frac{1}{8}(1-e^{-\frac{\mu}{2}\tau_3})^2(1-e^{-\frac{\mu}{2}(\tau_3+\tau_2)})+\frac{1}{8}(1+e^{-\frac{\mu}{2}\tau_3})^2(1+e^{-\frac{\mu}{2}(\tau_3+\tau_2)}) \tag{73}$$

Plugging into (70) we obtain the expression for η. Substituting z, g, h and η into (68) and rearranging terms, we obtain

$$\sigma R+S>T+\sigma P \tag{74}$$

where

$$\sigma=\frac{1+\nu+\mu}{3+\nu+\mu}\cdot\frac{K^*(\nu^2+2\nu+\nu\mu)+M(3+2\nu+\mu)}{K^*(\nu^2+2\nu+\nu\mu)+M(1+\mu)} \tag{75}$$

Note that if $\nu\neq 0$, the condition $\sigma>1$ is equivalent to $M>K^*$, which is always true. Therefore, σ is always larger than one when $\nu\neq 0$. Furthermore, σ is exactly one in the following limits: when $\nu\to 0$, when $\nu\to\infty$ or when $M/K^*\to 1$. In all of these cases, the population is well-mixed.

For $\mu=0$, we obtain

$$\sigma=\frac{1+\nu}{3+\nu}\cdot\frac{K^*\nu(2+\nu)+M(3+2\nu)}{K^*\nu(2+\nu)+M} \tag{76}$$

We observe that σ is a one-humped function of ν. It attains a maximum, which we denote by σ_{max}. To find the value of ν for which this maximum is achieved we differentiate (76) and set it equal to zero. We obtain the same expression as in (58). Therefore, it has the same solution which satisfies $\sqrt{M/K^*}<\nu_{opt}<\sqrt{M/K^*}+1$. For large M/K^*, the optimum ν is $\sqrt{M/K^*}$. Then we can write

$$\sigma_{max}=\frac{(1+\sqrt{M/K^*})^2}{3+\sqrt{M/K^*}} \tag{77}$$

Thus, when M/K^* is large, σ_{max} grows like $\sqrt{M/K^*}$.

We can also calculate σ_{max} for a fixed number of sets M. Since K^* is a function of L, K and M, each case has to be studied separately. For instance, if $L = K = 1$, then $K^* = 1$ and thus σ_{max} is proportional to $\sqrt{M}$ for M large enough.

If $K \geq 1$ and $L = K = M/2$ we find $K^* = M/\left(2\binom{M-1}{M/2-1}\right)$. When M is large, M/K^* is also large and thus σ_{max} grows proportional to $\sqrt{2\binom{M-1}{M/2-1}} \sim 2^{M/2}/M^{1/4}$. Thus, σ_{max} grows exponentially as a function of the number of sets, M.

Figure S5 shows the dependence of σ_{max} on M. The decisive condition for strategy A to be favored over B is $\sigma R + S > T + \sigma P$ (see eqn (74)). For a well-mixed population we have $\sigma = 1$. Larger values of σ indicate greater deviation from the well-mixed case and greater effect of the population structure. For $\sigma = 1$, strategy A is selected if $R + S > T + P$. In a coordination game ($R > T$ and $S < P$) this is the well-known condition of risk-dominance. For large values of σ, strategy A is selected if $R > P$. In a coordination game, this is the well-known condition of Pareto efficiency. Therefore evolutionary dynamics in set structured populations help to achieve the efficient equilibrium.

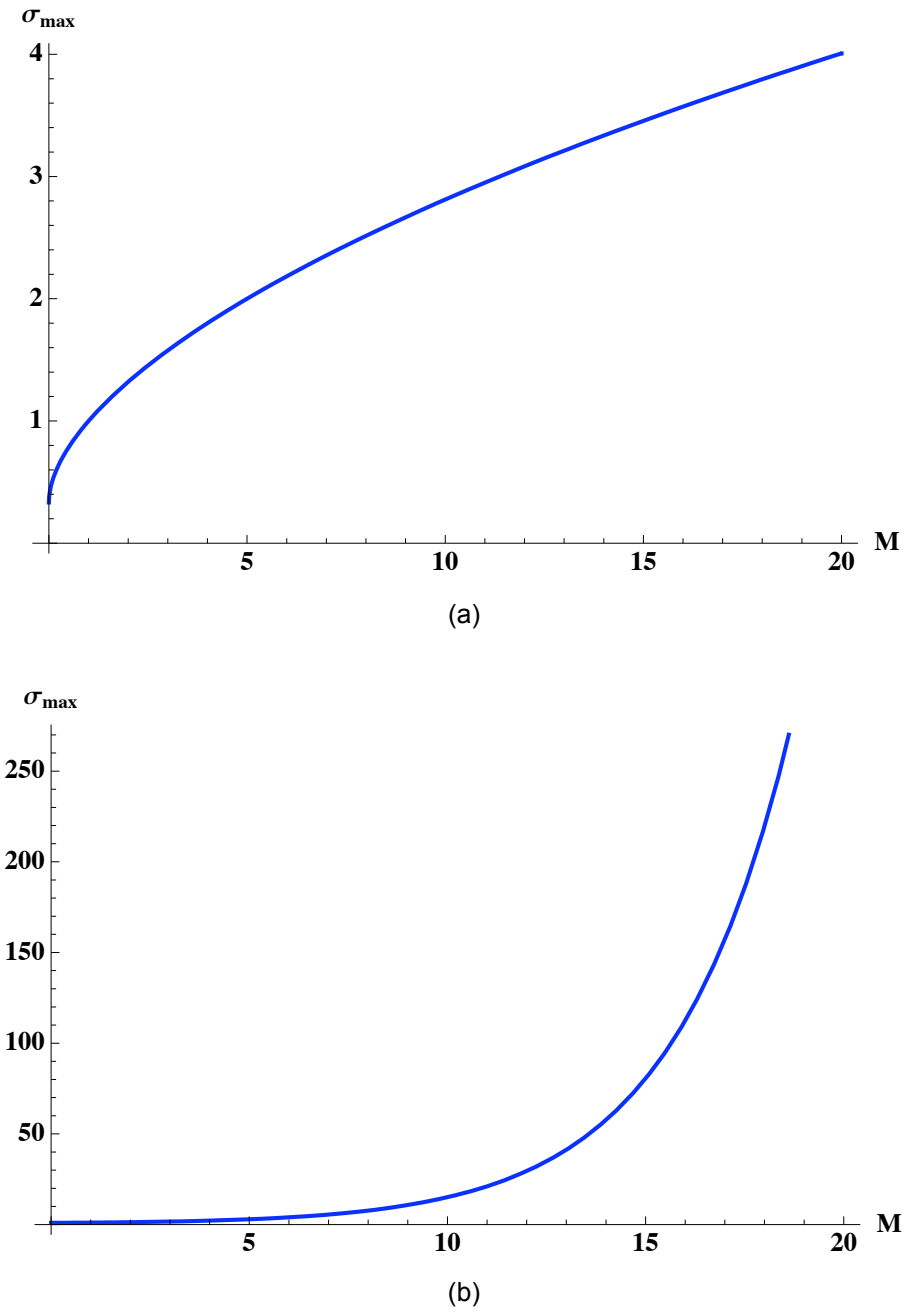


Fig. S 5: σ_{max} as a function of the number of sets M. We define σ_{max} to be the maximum value of σ given by eqn (76). **(a)** $L = K = 1$. **(b)** $L = K = M/2$.

Appendix – Analytic proof

Let ω_i represent the average number of offspring of individual i. After one update step (which is one generation) we have:

$$\omega_i = \frac{N f_i}{\sum_j f_j} \tag{78}$$

An individual is chosen to be a parent with probability given by its payoff relative to the total payoff and this choice is made independently N times per update step. Using (2) we can rewrite the denominator of ω_i as:

$$\sum_j f_j = N + \delta(b-c) \sum_j s_j \left(h_j \cdot \sum_i h_i - K \right) \tag{79}$$

We are interested in the limit of weak selection, $\delta \to 0$. Then, (78) becomes

$$\omega_i = 1 + \delta \left[-c s_i \left(h_i \cdot \sum_j h_j - K \right) + b h_i \cdot \sum_{j \neq i} s_j h_j - \frac{b-c}{N} \sum_{i,j,l} s_i s_j h_j \cdot h_l \right] + \mathcal{O}(\delta^2) \tag{80}$$

Let p denote the frequency of cooperators in the population. We think about an update step as having two parts: a selection part and a mutation part. We want to study the effect of mutation and selection on the average change in p. We denote by $\langle \Delta p \rangle_{\text{sel}}$ the effect due to selection, and by $\langle \Delta p \rangle_{\text{mut}}$ the effect due to mutation. Since the average value of p is constant, these effects must cancel:

$$\langle \Delta p \rangle_{\text{sel}} + \langle \Delta p \rangle_{\text{mut}} = 0 \tag{81}$$

In what follows, we will show that $\langle \Delta p \rangle_{\text{sel}}$ is continuous as a function of δ. Then, we can write its Taylor expansion at $\delta = 0$ using the fact that when $\delta = 0$ both of the above terms go to zero due to the symmetry of strategies in the neutral state

$$\langle \Delta p \rangle_{\text{sel}} = 0 + \delta \langle \Delta p \rangle^{(1)}_{\text{sel}} + \mathcal{O}(\delta^2) \tag{82}$$

Here $\langle \Delta p \rangle^{(1)}_{\text{sel}}$ is the first derivative of $\langle \Delta p \rangle_{\text{sel}}$ with respect to δ, evaluated at $\delta = 0$.

Thus, when $\langle \Delta p \rangle^{(1)}_{\text{sel}}$ is positive, it means that there is an increase in the frequency of cooperators due to selection. In this case, selection favors cooperators. If it is negative, then selection favors

defectors. Therefore, for the critical parameter values we must have

$$\langle \Delta p \rangle_{\text{sel}}^{(1)} = 0 \tag{83}$$

This condition holds for arbitrary values of the mutation rates. As the mutation rate goes to zero, the above equation corresponds to the equality of the fixation probabilities.

We will next detail the calculation of the change due to selection which we can write as

$$\langle \Delta p \rangle_{\text{sel}} = \sum_S (\Delta p)_S \cdot \pi_S \tag{84}$$

Here $(\Delta p)_S$ is the average number of A individuals in a given state S of the system and π_S is the probability to find the system in that state. Next we detail the calculation of this quantity.

Let s_i be the strategy of individual i, where $s_i = 1$ denotes A and $s_i = 0$ denotes B. Then, in a given state S of the system, the expected change of p due to selection in one update step is the number of offspring of A individuals minus the number of A individuals in the previous generation, divided by the population size. Thus, it is given by

$$(\Delta p)_S = \frac{1}{N} \left(\sum_i s_i \omega_i - \sum_i s_i \right) \tag{85}$$

where ω_i is the expected number of offspring of individual i.

From (80), we see that ω_i is a polynomial in δ. Hence, $(\Delta p)_S$ is also a polynomial in δ and thus it is continuous and infinitely differentiable at $\delta = 0$. Hence, we can write the Taylor expansion, using the fact that for the ω_i function (80), $(\Delta p)_S^{(0)} = 0$

$$(\Delta p)_S = 0 + \delta \left. \frac{\mathrm{d}(\Delta p)_S}{\mathrm{d}\delta} \right|_{\delta=0} + \mathcal{O}(\delta^2) = \frac{\delta}{N} \sum_i s_i \left. \frac{\mathrm{d}\omega_i}{\mathrm{d}\delta} \right|_{\delta=0} + \mathcal{O}(\delta^2) \tag{86}$$

The probability π_S that the system is in state S is also a function of δ. We will show that π_S is continuous and infinitely differentiable around $\delta = 0$ and thus that we can write its Taylor expansion

$$\pi_S = \pi_S^{(0)} + \delta \pi_S^{(1)} + \mathcal{O}(\delta^2) \tag{87}$$

The 0 superscript refers to the neutral state, $\delta = 0$ and $\pi_S^{(1)}$ is the first derivative of π_S as a function

of δ, evaluated at $\delta = 0$.

Next we show that π_S is continuous at $\delta = 0$ for all S. In order to find π_S, we need the transition probabilities P_{ij} to go from state S_j to state S_i. Then the stationary distribution is given by a vector of probabilities π_S, which is a normalized eigenvector corresponding to eigenvalue 1 of the stochastic matrix P. In our case, for the Wright-Fisher process with mutation, there is no absorbing subset of states. From every state of the system you can eventually reach every other state. This means that the matrix P is primitive, i.e. there exists some integer k such that $P^k > 0$.

For a primitive, stochastic matrix P, the Perron-Frobenius theorem ensures that 1 is its largest eigenvalue, that it is a simple eigenvalue and that it has a corresponding unique eigenvector with positive entries summing up to 1. This is precisely our vector of probabilities which gives the stationary distribution.

To find this eigenvector we perform Gaussian elimination (also referred to as row echelon reduction) on the system $Pv = v$. Since 1 is a simple eigenvalue for P, the system we need to solve has only one degree of freedom; thus we can express the eigenvector in terms of the one free variable, which without loss of generality can be v_n:

$$v_1 = -v_n a_1, \quad \dots \quad v_i = -v_n a_i, \quad \dots \quad v_{n-1} = -v_n a_{n-1} \tag{88}$$

The eigenvector that we are interested in is the vector with non-zero entries which sum up to 1. For this vector we have

$$1 = v_n(-a_1 - \ldots - a_{n-1} + 1) \tag{89}$$

This proof is general for any primitive stochastic matrix. Let us now return to our structure and the WF process. In our case, the transition probabilities come from the fitness $f = 1+\delta\cdot$payoff; they are fractions of such expressions and thus they are continuous at $\delta = 0$ and have Taylor expansions around $\delta = 0$. Thus, we can write all transition probabilities as polynomials in δ. Because of the elementary nature of the row operations performed, the elements of the reduced matrix are fractions of polynomials (i.e. rational functions of δ). Thus, a_i above are all rational functions of δ. Therefore, from (89) we conclude that v_n must also be a rational function of δ. This implies that in our vector of probabilities, all the entries are rational functions.

Thus π_S is a fraction of polynomials in δ which we write in irreducible form. The only way

that this is not continuous at $\delta = 0$ is if the denominator is zero at $\delta = 0$. But in that case, $\lim_{\delta\to 0} \pi_S = \infty$ which is impossible since π_S is a probability. Therefore, π_S is continuous at $\delta = 0$.

Once we have the Taylor expansions for both $(\Delta p)_S$ and π_S we can substitute them into (84) to obtain

$$\langle \Delta p \rangle_{\text{sel}} = \sum_S (\Delta p)_S \cdot \pi_S = \delta \sum_S \left. \frac{\mathrm{d}(\Delta p)_S}{\mathrm{d}\delta} \right|_{\delta=0} \cdot \pi_S^{(0)} + \mathcal{O}(\delta^2) \tag{90}$$

$$= \frac{\delta}{N} \sum_S \left(\sum_i s_i \left. \frac{\mathrm{d}\omega_i}{\mathrm{d}\delta} \right|_{\delta=0} \right) \cdot \pi_S^{(0)} + \mathcal{O}(\delta^2) \tag{91}$$

$$=: \frac{\delta}{N} \left\langle \sum_i s_i \frac{\mathrm{d}\omega_i}{\mathrm{d}\delta} \right\rangle_0 + \mathcal{O}(\delta^2) \tag{92}$$

The last line is just notation. The angular brackets denote the average and the 0 subscript refers to the neutral state $\delta = 0$. Note that we start by writing the average change in the presence of the game in equation (90) and we end up with an expression depending on the neutral state (92), but containing the parameters b and c. Therefore we have shown that we only need to do our calculations in the neutral state.

Now using (80) in (92), the first derivative of the effect of selection in the stationary state evaluated at $\delta = 0$ becomes

$$\begin{aligned} \langle \Delta p \rangle_{\text{sel}}^{(1)} = \frac{1}{N} \Bigg[& -c \left\langle \sum_{i,j} s_i \gamma_{ij} h_i \cdot h_j \right\rangle_0 - K(b-c) \left\langle \sum_i s_i \right\rangle_0 + K \frac{b-c}{N} \left\langle \sum_{i,j} s_i s_j \right\rangle_0 \\ & + b \left\langle \sum_{i,j} s_i s_j \gamma_{ij} h_i \cdot h_j \right\rangle_0 + \frac{b-c}{N} \left\langle \sum_{i,j,l} s_i s_j \gamma_{jl} h_j \cdot h_l \right\rangle_0 \Bigg] \end{aligned} \tag{93}$$

As discussed above, the critical b/c ratio is obtained when equation (83) holds. From this we obtain

$$\left(\frac{b}{c}\right)^* = \frac{\langle \sum_{i,j} s_i \gamma_{ij} h_i \cdot h_j \rangle_0 - \frac{1}{N} \langle \sum_{i,j,l} s_i s_j \gamma_{jl} h_j \cdot h_l \rangle_0 - K \langle \sum_i s_i \rangle_0 + \frac{K}{N} \langle \sum_{i,j} s_i s_j \rangle_0}{\langle \sum_{i,j} s_i s_j \gamma_{ij} h_i \cdot h_j \rangle_0 - \frac{1}{N} \langle \sum_{i,j,l} s_i s_j \gamma_{jl} h_j \cdot h_l \rangle_0 - K \langle \sum_i s_i \rangle_0 + \frac{K}{N} \langle \sum_{i,j} s_i s_j \rangle_0} \tag{94}$$

Hence, we have derived the expression for the critical b/c ratio given by (12).

References

[1] Fisher RA (1930) The Genetical Theory of Natural Selection. (Oxford: Clarendon Press).

[2] Wright S (1931) Evolution in mendelian populations. *Genetics* 16: 97-159.

[3] Ewens WJ (2004) Mathematical population genetics. Theoretical introduction, vol. 1. (Springer, New York).

[4] Maynard Smith J (1982) Evolution and the Theory of Games. (Cambridge Univ. Press, Cambridge, UK).

[5] Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics. (Cambridge Univ. Press, Cambridge, UK).

[6] Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428: 646-650.

[7] Antal T, Ohtsuki H, Wakeley J, Taylor PD, Nowak MA (2008) Evolutionary game dynamics in phenotype space, e-print arXiv:0806.2636.

[8] Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359: 826-829.

[9] Ohtsuki H, Hauert C, Lieberman E, Nowak MA (2006) A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441: 502-505.

[10] Ohtsuki H, Nowak MA (2008) Evolutionary stability on graphs. *J theor Biol* 251: 698-707.